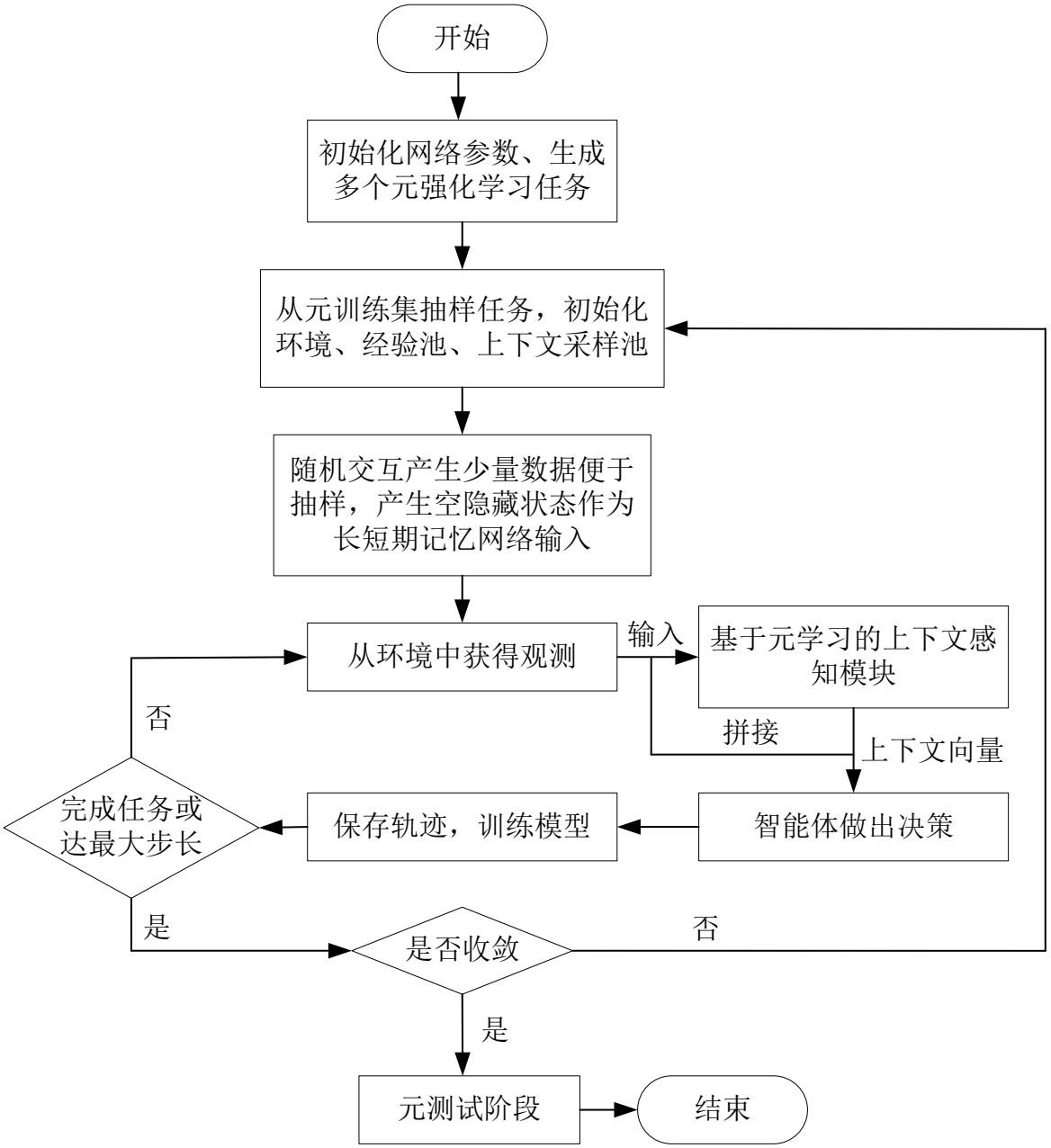
说 明 书 摘 要

本发明公开了一种基于元强化学习的无人集群抗干扰通信方法，具体为：针对无人集群网络复杂电磁环境中的未知动态干扰，在元训练流程中，各无人机首先将与环境交互的观测数据输入至用于上下文任务推理的编码器和长短期记忆网络模块，从编码器输出的分布上进行抽样得到上下文任务信息，长短期记忆网络的隐藏状态作为当前任务的历史记忆，再将其与智能体的当前观测拼接以作为双深度 Q 网络（DDQN）输入，预测动作并与环境交互。通过生成潜在上下文变量的先验分布，能推断如何从少量经验中解决新任务，该方法具有可迁移性强、样本效率高和抗干扰性能优越的优点，能够快速适应不同任务，在实现高效的数据传输的同时最小化传输能量损耗和跳频开销。

摘 要 附 图


开始
初始化网络参数、生成多个元强化学习任务
从元训练集抽样任务，初始化环境、经验池、上下文采样池
随机交互产生少量数据便于抽样，产生空隐藏状态作为长短期记忆网络输入
从环境中获得观测
输入
基于元学习的上下文感知模块
拼接
上下文向量
智能体做出决策
保存轨迹，训练模型
完成任务或达最大步长
否
是
是否收敛
否
是
元测试阶段
结束

1. 一种基于元强化学习的无人集群抗干扰通信技术，其特征在于，包括如下步骤：

步骤 1：构建多智能体无人集群抗干扰通信环境并定义系统环境状态空间和动作空间，初始化多智能体各训练网络权重，初始化多智能体经验池和上下文采样池，初始化基于上下文任务推理的编码器和长短期记忆网络模型，根据所定义环境生成多个元强化学习任务；

步骤 2：各智能体随机与环境交互采集少量数据存于经验池和采样池中，并产生一个空的隐藏状态；

步骤 3：智能体通过计算得到相应动作，与环境交互并获得相应奖励；

步骤 4：更新上下文采样器和经验池；

步骤 5：计算损失函数，并反向传播更新各网络参数；

步骤 6：若环境达到最大训练步长，结束当前回合，重置多智能体无人集群抗干扰环境，初始化经验池和采样池，进入元测试阶段，否则重复步骤 2 开始下一轮的训练。

2. 根据权利要求 1 所述的基于元强化学习的无人集群抗干扰通信技术，其特征在于，步骤 1 中所述的多智能体抗干扰通信环境包含：

(1)网络模型：包含 N 个干扰机和 M 个无人机集群，其中每个无人机集群包含一个簇头和 I 个簇成员，干扰机与簇头按照马尔可夫随机移动模型移动，簇成员按照参考点组移动模型，参考点在以簇头无人机为中心，半径为 R_1 的圆形区域内移动，簇成员以各自参考点为中心，在半径为 R_2 的圆形区域内移动，各参考点的移动半径与簇内成员的移动半径之和不超过最大通信距离 $D_{\max}$，即满足 $R_1 + R_2 \le D_{\max}$。

(2)干扰模型：干扰机使用马尔可夫干扰模式，它在信道转换期间遵循马尔可夫性质，从信道 l_i^{t-1} 转移到信道 $[l_1^t, l_2^t, ..., l_L^t]$ 的概率记为 $\mathbf{pr}^t \triangleq [pr_1^t, pr_2^t, ..., pr_L^t]$，$\mathbf{pr}^t$ 初始化为 $[0,1]$ 的均匀分布，其中 L 为信道总数。记操作 A：以均等的概率在 $[1,2,...,L]$ 中随机抽取一个信道 l_{random}^t，设 $\text{Sum}_{pr} = \sum_i^L pr_i^t$，从信道 l_i^{t-1} 转移到信道 l_{random}^t 的概率加上 $\text{Sum}_{pr} / \text{div}$，其中 div 为一个倍数缩放因子。操作 B：持续生成 $[0,1]$ 之间的随机数，若该随机数大于 pr_{continue}，则执行一次操作 A，直至生成的随机数不大于 pr_{continue} 为止。操作 C：将 $\mathbf{pr}^t$ 归一化。训练任务利用操作 ABC 的组合生成，以获得多样的元强化学习任务。

(3)通信模型：考虑将时间划分为 T 个时隙，每个时隙的持续时间为 dt，功率增益由慢衰落和快衰落组成，设 $f_{i,m}^{t,l}$ 表示在时隙 t 信道 l 上快衰落增益，$\mathcal{N}$ 为标准正态分布，簇头无人机 m 和簇成员 i 之间的 $f_{i,m}^{t,l}$ 可以表示为:

$$f_{i,m}^{t,l}=20\times\log(\frac{1}{\sqrt{2}}\|\mathcal{N}(1,0)+j\times\mathcal{N}(1,0)\|)$$

慢衰落表示为$PL_{i,m}^{t,l}=A\times\mathrm{dis}(i,m)^{-\zeta}$，其中$A$为参考距离1m处的路径损耗常量，$\zeta$为路径损耗分量的衰减指数，$\mathrm{dis}(i,m)$表示无人机$i$和无人机$m$之间的距离。在时隙$t$，CH$m$与CM$i$在第$l$个信道上通信的信道功率增益可以表示为$g_{i,m}^{t,l}=f_{i,m}^{t,l}PL_{i,m}^{t,l}$。定义干扰机$i$在时隙$t$的链路增益信息为$\overline{g}_i^t$，传输功率为$\overline{p}_i^t$，干扰机信道选择为$\overline{\mathbf{L}}^t\triangleq[\overline{l}_1^t,...,\overline{l}_I^t]$。因此，在时段$t$，CH$m$和CM$i$之间的信道$l$的干扰可以表示为:

$$V_{i,m}^{t,l}=\sum_{q=1}^{M}\sum_{i=1}^{I}1_{\{l_{q,i}^t=l,q\neq m\}}p_{q,i}^t g_{q,i}^{t,l}+\sum_{n=1}^{N}1_{\{\overline{l}_n=l\}}\overline{p}_n^t\overline{g}_n^{t,l}+\sum_{k=1}^{I}1_{\{l_{m,k}^t=l,k\neq i\}}p_{m,k}^t g_{m,k}^{t,l}$$

令σ^2为高斯加性白噪声的功率，B为每个信道的带宽。则簇头无人机m与簇成员i在时隙t信道l上的数据传输速率为:

$$R_{i,m}^{t,l}=B\log_2\left(1+\frac{p_{i,m}^t g_{i,m}^{t,l}}{\sigma^2+V_{i,m}^{t,l}}\right)$$

进一步地，步骤 1 中所述的基于元强化学习的无人集群抗干扰通信技术具体将优化问题建模为一个部分可观测的马尔可夫决策过程（POMDP），智能体m在时隙t的观测值可以定义为$o_m^t\triangleq[\mathbf{G}_m^t,\overline{\mathbf{L}}^t]$，其中$\mathbf{G}_m^t\triangleq[\mathbf{G}_{m,1}^t,...,\mathbf{G}_{m,I}^t]$表示在时隙$t$中，簇头无人机$m$与其簇内所有成员在所有通信信道上的信道功率增益，智能体的动作表示为$a_m^t\triangleq[\mathbf{L}_m^t,\mathbf{P}_m^t]$。其中，$\mathbf{L}_m^t$表示传输信道的选择，$\mathbf{P}_m^t$表示发射功率，智能体$m$在$o_m^t$的基础上选择动作$a_m^t$的策略，可以表示为$\pi_m(a_m^t\mid o_m^t)\triangleq\Pr(a_m^t\mid o_m^t)$，Pr表示概率。设传输能量损耗$E_m^t=\sum_{i=1}^{I}\sum_{l=1}^{L}1\{l_{m,i}^t=l\}p_{m,i}^t T_{m,i}^{t,l}$和跳频开销$F_m^t=\sum_{i=1}^{I}\sum_{l=1}^{L}1\{l_{m,i}^t\neq l_{m,i}^{t-1}\}W$，$\beta$调节$E_{m,i}^t$和$F_m^t$的比重，$W$为单位跳频成本。考虑到干扰机在数据传输阶段$T_{Rx}$内可能切换干扰信道，将传输时间$T_{Rx}$分为两个时间段$T_1$和$T_2$，满足$T_1+T_2=T_{Rx}$.假设每个时隙传输的数据包大小为$K$，则在时隙$t$中，簇头无人机$m$与其簇内成员$i$在信道$l$上实际传输时间为：

$$T_{m,i}^{t,l}=\begin{cases}\dfrac{K}{R1_{m,i}^{t,l}} & ,若\dfrac{K}{R1_{m,i}^{t,l}}<T_1,\\ T_{Rx} & ,其他,\\ T_1+\dfrac{K-R1_{m,i}^{t,l}T_1}{R2_{m,i}^{t,l}} & ,若\dfrac{K}{R1_{m,i}^{t,l}}>T_1且\dfrac{K-R1_{m,i}^{t,l}T_1}{R2_{m,i}^{t,l}}<T_2,\end{cases}$$

其中，$R1_{m,i}^{t,l}$和$R2_{m,i}^{t,l}$分别表示时隙t中T_1和T_2时间段内在信道l上簇头无人机m与其簇内成员i的数据传输速率，设辅助变量$Z_m^t=\sum_{i=1}^{I}1\{R1_{m,i}^{t,l}T_1+R2_{m,i}^{t,l}T_2>K\}$，则智能体$m$通过执行动作$a_m^t$会获得的本地奖励$r_m^t=Z_m^t\delta-[\beta E_m^t+(1-\beta)F_m^t]$，$\delta$为数据传输成功奖励。

3. 根据权利要求 1 所述的基于元强化学习的无人集群抗干扰通信技术，其特征在于，步

骤 2 和步骤 4 中经验池和采样器将任务Γ中的一个转移$c_m^{\Gamma,t}=(o_m^t,o_m^{t+1},a_m^t,r_m^t)$存入经验池$\mathcal{D}_m$和上下文采样池$\mathcal{C}_m$，每个簇头$m$的经验池和上下文采样池相互独立。

4.根据权利要求 1 所述的基于元强化学习的无人集群抗干扰通信技术，其特征在于，步骤 3 中描述获得相应动作的过程为：基于 DDQN 算法新增上下文元强化学习框架，从上下文采样器中抽样多组数据$\mathbf{c}_m^{\Gamma}\sim\mathcal{C}_m$作为上下文任务推理编码器的输入，将隐藏状态$h^{t-1}$和环境当前观测值$o^t$作为长短期记忆网络的输入，将从上下文任务推理编码器得到的向量z、长短期记忆网络的输出h^t、环境当前观测值连接作为 DDQN 的输入。具体的，用推理网络$q_\phi(z|\mathbf{c}^{\Gamma})$用于估计后验分布$p(z|\mathbf{c}^{\Gamma})$，由马尔可夫决策过程（MDP）的置换不变性将$q_\Phi(z|c_{1:N})$建模为：

$$q_\phi(z|c_{1:N})\propto\prod_{n=1}^{N}\Psi_\phi(z|c_n)$$

其中$\Psi_\phi(z|c_n)=\mathrm{N}\left(f_\phi^\mu(c_n),f_\phi^\sigma(c_n)\right)$，它是先由$f_\phi^\mu(c_n)$和$f_\phi^\sigma(c_n)$的神经网络估计初均值$\mu$和方差$\sigma$，然后再从正态分布中进行随机抽样得来。LSTM 用于对序列信息进行处理，在遗忘门中，$f^t=\sigma(W_f[h^{t-1},o^t]+b_f)$，其中的$h^{t-1}$为$t-1$时刻的隐藏状态，$o^t$为$t$时刻的状态。

5.根据权利要求 1 所述的基于元强化学习的无人集群抗干扰通信技术，其特征在于，步骤 5 中描述的更新方法为：令$G_t=r^t+\gamma r^{t+1}+.....+\gamma^k r^{t+k}$，在 DDQN 评估网络中$Q_\pi(o^t,a^t;\mathbf{w}^t)=E[G_t|o^t,a^t;\mathbf{w}^t]$用于评估在网络参数为$\mathbf{w}^t$时当前观测$o^t$和动作$a^t$的$Q_{\text{eval}}$值，目标网络$\hat{Q}_\pi$用于评估下个时刻的$Q_{\text{target}}$值，定义损失函数$\mathcal{L}=(r^t+\gamma\tilde{Q}_{\text{target}}-\tilde{Q}_{\text{eval}})^2$，其中$\tilde{Q}_{\text{target}}$和$\tilde{Q}_{\text{eval}}$为当前观测下最大的$Q$值，即选择动作的$Q$值，$Q_\pi$的参数$\mathbf{w}$使用 Adam 算法进行优化，$\hat{Q}_\pi$的参数每隔一段时间从$Q_\pi$参数复制。

6. 根据权利要求 1 所述的基于元强化学习的无人集群抗干扰通信技术，其特征在于，步骤 6 中描述的元测试阶段具体过程为：利用预先训练好的模型参数在元测试集上进行测试，在测试过程对模型参数进行微调，测试其快速适应新的环境的能力和最终收敛效果。

一种基于元强化学习的无人集群抗干扰通信方法

技术领域

本发明属于无人机智能通信技术领域，特别是涉及一种基于元强化学习的无人集群抗干扰通信方法。

背景技术

无人机因其高可靠性、灵活性和适中的价格，作为一种新技术备受关注。然而，在复杂的电磁环境中，无人机集群网络通信面临恶意干扰信号和频谱资源稀缺两大挑战，恶意干扰信号会破坏无线电波的传播，而有限的频谱资源则使得不同集群之间的通信链路容易受到同频干扰，传统的抗干扰通信方案无法适应复杂电磁环境中动态变化的干扰模式。此外，面对有限的电池容量和频谱资源，需要有效实施功率控制以延长网络寿命。因此，如何构建具有强大抗干扰能力和高传输可靠性的长寿命无人机集群网络，已成为当前无人机集群领域亟待解决的关键问题之一。

基于认知无线电的概念，可以实现动态频谱抗干扰通信技术，以提升无人机集群网络的性能。例如，无线系统应用 Q 学习[Wang X, Jin T, Hu L, et al. Energy-efficient power allocation and Q-learning-based relay selection for relay-aided D2D communication[J]. IEEE Transactions on Vehicular Technology, 2020, 69(6): 6452-6462.]，根据网络拓扑和信道条件选择中继节点向相邻无线电设备传输数据，该系统还利用拉格朗日对偶分解法分配传输功率，提高能源效率和抗干扰能力。此外，另一种值得关注的方案是一种基于多智能体强化学习的无人机集群通信方法[Lv Z, Niu G, Xiao L, et al. Reinforcement learning based UAV swarm communications against jamming[C]. ICC 2023-IEEE International Conference on Communications. IEEE, 2023: 5204-5209.]，通过优化无人机中继选择和功率分配来增强通信性能并节省能源消耗。最近，还有一项研究提出了利用双深度 Q 学习方法，针对联合频谱域和功率域的无人机簇群抗干扰方法[Wu Z,Lin Y,Zhang Y,et al. Multi-domain energy-saving and anti-jamming communication of UAV cluster based on multi-agent cooperation [J]. Science China: Information Science, 2023(12):2511-2526.]，相较于基准算法能更有效地降低长期传输能量损耗和跳频开销，且同时提升数据传输成功率。然而，环境和任务的动态性给决策带来了新的挑战，这些传统的强化学习方法需要对每个新任务重新训练模型，无法充分应对环境和任务的变化。

元强化学习作为近年来强化学习研究的一个重要分支，受到了越来越多的关注。元强化学习旨在训练一个只需少量交互就能快速学习和适应新任务的模型，传统强化学习需要为每个

新任务从头开始重新学习。例如，Finn 等人提出的模型不可知元学习(Model Agnostic Meta-Learning, MAML)[Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks[C]. International Conference on Machine Learning. PMLR, 2017: 1126-1135.]是一种基于优化元目标的元学习方法，它跨任务训练一组共享的元参数θ，这样为新任务初始化的模型只需要少量数据就可以快速收敛。Duan 等人提出了一种RL^2算法[Duan Y, Schulman J, Chen X, et al. RL^2: Fast reinforcement learning via slow reinforcement learning[J]. arXiv preprint arXiv:1611.02779, 2016.]，该算法使用递归神经网络构建策略模型，并跨多个任务进行训练，在相同任务的迭代之间传递隐藏状态，在不同任务之间传递参数，目的是利用内存优化跨任务能力。Rakelly 等人提出了 Probabilistic Embeddings for Actor-critic meta-RL (PEARL)算法[Rakelly K, Zhou A, Finn C, et al. Efficient off-policy meta-reinforcement learning via probabilistic context variables[C]. International conference on machine learning. PMLR, 2019: 5331-5340.]，该算法以任务历史信息为输入，要求智能体推断当前任务的特征，从而将部分可观察马尔可夫决策过程任务映射到马尔可夫决策过程中来构建策略。

综上所述，无人机通信技术面临的挑战和复杂性显而易见，尽管有一些强化学习方法已经被提出并应用于这一领域，但在处理大规模、动态变化的环境以及复杂的干扰模式时，仍然存在许多挑战，元强化学习体现了在处理这些问题上的潜力和优势，是一种更高级的学习方法，它试图学习如何快速适应新的任务，以便在少量的数据中取得最大的学习效果，这使得元强化学习在处理无人机通信技术中的干扰问题时具有独特的优势。

发明内容

本发明旨在提供一种基于元强化学习的无人集群抗干扰通信技术，能够稳健地增强模型性能，使得模型能够更快速地学习适应新任务，使得模型在面对新任务时表现更加出色。

实现本发明的技术解决方案为：将一种基于上下文感知的 DDQN 元学习抗干扰算法用于无人集群抗干扰通信，具体步骤为：

步骤 1：构建多智能体无人集群抗干扰通信环境并定义系统环境状态空间和动作空间，初始化多智能体各训练网络权重，初始化多智能体经验池和上下文采样池，初始化基于上下文任务推理的编码器和长短期记忆网络模型，根据所定义环境生成多个元强化学习任务；

步骤 2：各智能体随机与环境交互采集少量数据存于经验池和采样池中，并产生一个空的隐藏状态；

步骤 3：智能体通过计算得到相应动作，与环境交互并获得相应奖励；

步骤 4：更新上下文采样器和经验池；

步骤 5：计算损失函数，并反向传播更新各网络参数；

步骤 6：若环境达到最大训练步长，结束当前回合，重置多智能体无人集群抗干扰环境，初始化经验池和采样池，进入元测试阶段，否则重复步骤 2 开始下一轮的训练。

进一步地，步骤 1 中所述的多智能体抗干扰通信环境包含：

(1)网络模型：包含 N 个干扰机和 M 个无人机集群，其中每个无人机集群包含一个簇头和 I 个簇成员，干扰机与簇头按照马尔可夫随机移动模型移动，簇成员按照参考点组移动模型，参考点在以簇头无人机为中心，半径为 R_1 的圆形区域内移动，簇成员以各自参考点为中心，在半径为 R_2 的圆形区域内移动，各参考点的移动半径与簇内成员的移动半径之和不超过最大通信距离 $D_{\max}$，即满足 $R_1+R_2\le D_{\max}$。

(2)干扰模型：干扰机使用马尔可夫干扰模式，它在信道转换期间遵循马尔可夫性质，从信道 l_i^{t-1} 转移到信道 $[l_1^t,l_2^t,...,l_L^t]$ 的概率记为 $\mathbf{pr}^t\triangleq[pr_1^t,pr_2^t,...,pr_L^t]$，$\mathbf{pr}^t$ 初始化为 $[0,1]$ 的均匀分布，其中 L 为信道总数。记操作 A：以均等的概率在 $[1,2,...,L]$ 中随机抽取一个信道 l_{random}^t，设 $\text{Sum}_{pr}=\sum_i^L pr_i^t$，从信道 l_i^{t-1} 转移到信道 l_{random}^t 的概率加上 $\text{Sum}_{pr}/\text{div}$，其中 div 为一个倍数缩放因子。操作 B：持续生成 $[0,1]$ 之间的随机数，若该随机数大于 pr_{continue}，则执行一次操作 A，直至生成的随机数不大于 pr_{continue} 为止。操作 C：将 $\mathbf{pr}^t$ 归一化。训练任务利用操作 ABC 的组合生成，以获得多样的元强化学习任务。

(3)通信模型：考虑将时间划分为 T 个时隙，每个时隙的持续时间为 dt，功率增益由慢衰落和快衰落组成，设 $f_{i,m}^{t,l}$ 表示在时隙 t 信道 l 上快衰落增益，$\mathcal{N}$ 为标准正态分布，簇头无人机 m 和簇成员 i 之间的 $f_{i,m}^{t,l}$ 可以表示为:

$$f_{i,m}^{t,l}=20\times\log(\frac{1}{\sqrt{2}}\|\mathcal{N}(1,0)+j\times\mathcal{N}(1,0)\|)$$

慢衰落表示为 $PL_{i,m}^{t,l}=A\times\text{dis}(i,m)^{-\zeta}$，其中 A 为参考距离 1m 处的路径损耗常量，ζ 为路径损耗分量的衰减指数，$\text{dis}(i,m)$ 表示无人机 i 和无人机 m 之间的距离。在时隙 t，CH m 与 CM i 在第 l 个信道上通信的信道功率增益可以表示为 $g_{i,m}^{t,l}=f_{i,m}^{t,l}PL_{i,m}^{t,l}$。定义干扰机 i 在时隙 t 的链路增益信息为 $\overline{g}_i^t$，传输功率为 $\overline{p}_i^t$，干扰机信道选择为 $\overline{\mathbf{L}}^t\triangleq[\overline{l}_1^t,...,\overline{l}_I^t]$。因此，在时段 t，CH m 和 CM i 之间的信道 l 的干扰可以表示为:

$$V_{i,m}^{t,l}=\sum_{q=1}^{M}\sum_{i=1}^{I}1_{\{l_{q,i}^t=l,q\neq m\}}p_{q,i}^t g_{q,i}^{t,l}+\sum_{n=1}^{N}1_{\{\overline{l}_n=l\}}\overline{p}_n^t\overline{g}_n^{t,l}+\sum_{k=1}^{I}1_{\{l_{m,k}^t=l,k\neq i\}}p_{m,k}^t g_{m,k}^{t,l}$$

令 σ^2 为高斯加性白噪声的功率，B 为每个信道的带宽。则簇头无人机 m 与簇成员 i 在时隙 t 信

道 l 上的数据传输速率为:

$$R_{i,m}^{t,l}=B\log_2\left(1+\frac{p_{i,m}^{t}g_{i,m}^{t,l}}{\sigma^2+V_{i,m}^{t,l}}\right)$$

进一步地，步骤 1 中所述的基于元强化学习的无人集群抗干扰通信技术具体将优化问题建模为一个部分可观测的马尔可夫决策过程（POMDP），智能体 m 在时隙 t 的观测值可以定义为 $o_m^t \triangleq[\mathbf{G}_m^t,\bar{\mathbf{L}}^t]$，其中 $\mathbf{G}_m^t \triangleq[\mathbf{G}_{m,1}^t,...,\mathbf{G}_{m,I}^t]$ 表示在时隙 t 中，簇头无人机 m 与其簇内所有成员在所有通信信道上的信道功率增益，智能体的动作表示为 $a_m^t \triangleq[\mathbf{L}_m^t,\mathbf{P}_m^t]$。其中，$\mathbf{L}_m^t$ 表示传输信道的选择，$\mathbf{P}_m^t$ 表示发射功率，智能体 m 在 o_m^t 的基础上选择动作 a_m^t 的策略，可以表示为 $\pi_m(a_m^t\mid o_m^t)\triangleq\Pr(a_m^t\mid o_m^t)$，Pr 表示概率。设传输能量损耗 $E_m^t=\sum_{i=1}^{I}\sum_{l=1}^{L}1\{l_{m,i}^t=l\}p_{m,i}^tT_{m,i}^{t,l}$ 和跳频开销 $F_m^t=\sum_{i=1}^{I}\sum_{l=1}^{L}1\{l_{m,i}^t\neq l_{m,i}^{t-1}\}W$，$\beta$ 调节 $E_{m,i}^t$ 和 F_m^t 的比重，W 为单位跳频成本。考虑到干扰机在数据传输阶段 T_{Rx} 内可能切换干扰信道，将传输时间 T_{Rx} 分为两个时间段 T_1 和 T_2，满足 $T_1+T_2=T_{Rx}$.假设每个时隙传输的数据包大小为 K，则在时隙 t 中，簇头无人机 m 与其簇内成员 i 在信道 l 上实际传输时间为：

$$T_{m,i}^{t,l}=\begin{cases}\dfrac{K}{R1_{m,i}^{t,l}} & ,若\dfrac{K}{R1_{m,i}^{t,l}}<T_1,\\ T_{Rx} & ,其他,\\ T_1+\dfrac{K-R1_{m,i}^{t,l}T_1}{R2_{m,i}^{t,l}} & ,若\dfrac{K}{R1_{m,i}^{t,l}}>T_1且\dfrac{K-R1_{m,i}^{t,l}T_1}{R2_{m,i}^{t,l}}<T_2,\end{cases}$$

其中，$R1_{m,i}^{t,l}$ 和 $R2_{m,i}^{t,l}$ 分别表示时隙 t 中 T_1 和 T_2 时间段内在信道 l 上簇头无人机 m 与其簇内成员 i 的数据传输速率，设辅助变量 $Z_m^t=\sum_{i=1}^{I}1\{R1_{m,i}^{t,l}T_1+R2_{m,i}^{t,l}T_2>K\}$，则智能体 m 通过执行动作 a_m^t 会获得的本地奖励 $r_m^t=Z_m^t\delta-[\beta E_m^t+(1-\beta)F_m^t]$，$\delta$ 为数据传输成功奖励。

步骤 2 和步骤 4 中经验池和采样器将任务 Γ 中的一个转移 $c_m^{\Gamma,t}=(o_m^t,o_m^{t+1},a_m^t,r_m^t)$ 存入经验池 $\mathcal{D}_m$ 和上下文采样池 $\mathcal{C}_m$，每个簇头 m 的经验池和上下文采样池相互独立。

进一步地，步骤 3 中描述获得相应动作的过程为：基于 DDQN 算法新增上下文元强化学习框架，从上下文采样器中抽样多组数据 $\mathbf{c}_m^{\Gamma}\sim\mathcal{C}_m$ 作为上下文任务推理编码器的输入，将隐藏状态 h^{t-1} 和环境当前观测值 o^t 作为长短期记忆网络的输入，将从上下文任务推理编码器得到的向量 z、长短期记忆网络的输出 h^t、环境当前观测值连接作为 DDQN 的输入。具体的，用推理网络 $q_\phi(z\mid\mathbf{c}^{\Gamma})$ 用于估计后验分布 $p(z\mid\mathbf{c}^{\Gamma})$，由马尔可夫决策过程（MDP）的置换不变性将 $q_\Phi(z\mid c_{1:N})$ 建模为：

$$q_\phi(z\mid c_{1:N})\propto\prod\nolimits_{n=1}^{N}\Psi_\phi(z\mid c_n)$$

其中$\Psi_{\phi}(z \mid c_n)=\mathrm{N}\left(f_{\phi}^{\mu}(c_n), f_{\phi}^{\sigma}(c_n)\right)$，它是先由$f_{\phi}^{\mu}(c_n)$和$f_{\phi}^{\sigma}(c_n)$的神经网络估计初均值$\mu$和方差$\sigma$，然后再从正态分布中进行随机抽样得来。LSTM 用于对序列信息进行处理，在遗忘门中，$f^t=\sigma(W_f[h^{t-1},o^t]+b_f)$，其中的$h^{t-1}$为$t-1$时刻的隐藏状态，$o^t$为$t$时刻的状态。

进一步地，步骤 5 中描述的更新方法为：令$G_t=r^t+\gamma r^{t+1}+.....+\gamma^k r^{t+k}$，在 DDQN 评估网络中$Q_{\pi}(o^t,a^t;\mathbf{w}^t)=E[G_t \mid o^t,a^t;\mathbf{w}^t]$用于评估在网络参数为$\mathbf{w}^t$时当前观测$o^t$和动作$a^t$的$Q_{\text{eval}}$值，目标网络$\hat{Q}_{\pi}$用于评估下个时刻的$Q_{\text{target}}$值,定义损失函数$\mathcal{L}=(r^t+\gamma\tilde{Q}_{\text{target}}-\tilde{Q}_{\text{eval}})^2$，其中$\tilde{Q}_{\text{target}}$和$\tilde{Q}_{\text{eval}}$为当前观测下最大的$Q$值，即选择动作的$Q$值，$Q_{\pi}$的参数$\mathbf{w}$使用 Adam 算法进行优化，$\hat{Q}_{\pi}$的参数每隔一段时间从$Q_{\pi}$参数复制。

进一步地，步骤 6 中描述的元测试阶段具体过程为：利用预先训练好的模型参数在元测试集上进行测试，在测试过程对模型参数进行微调，测试其快速适应新的环境的能力和最终收敛效果。

本发明与现有技术相比，其显著优点为：（1）本发明高度模块化可扩展性强，能作为一个独立的模块被整合已有算法中而无需对原有系统进行大规模修改或重构。（2）本发明中的元学习机制展现出了更强的泛化性能，在面对新的干扰策略时，模型能够在更短的时间内收敛，显著提升了学习效率，并且能获得更好的收敛结果，这意味着模型在训练过程中能够更快地找到最优解，提高了整体的性能和稳定性。

附图说明

图 1 是元训练阶段累计奖励收敛效果示意图。

图 2 是元测试阶段累计奖励收敛效果示意图。

图 3 是元测试阶段累计能量奖励收敛效果示意图。

图 4 是元测试阶段累计跳频奖励收敛效果示意图。

图 5 是元测试阶段累计传输奖励收敛效果示意图。

具体实施方式

本发明旨在提供一种基于元强化学习的无人集群抗干扰通信技术，具体步骤为：

步骤 1：构建多智能体无人集群抗干扰通信环境并定义系统环境状态空间和动作空间，初始化多智能体各训练网络权重，初始化多智能体经验池和上下文采样池，初始化基于上下文任务推理的编码器和长短期记忆网络模型，根据所定义环境生成多个元强化学习任务；

步骤 2：各智能体随机与环境交互采集少量数据存于经验池和采样池中，并产生一个空的隐藏状态；

步骤 3：智能体通过计算得到相应动作，与环境交互并获得相应奖励；

步骤 4：更新上下文采样器和经验池；

步骤 5：计算损失函数，并反向传播更新各网络参数；

步骤 6：若环境达到最大训练步长，结束当前回合，重置多智能体无人集群抗干扰环境，初始化经验池和采样池，进入元测试阶段，否则重复步骤 2 开始下一轮的训练。

进一步地，步骤 1 中所述的多智能体抗干扰通信环境包含：

(1)网络模型：包含 N 个干扰机和 M 个无人机集群，其中每个无人机集群包含一个簇头和 I 个簇成员，干扰机与簇头按照马尔可夫随机移动模型移动，簇成员按照参考点组移动模型，参考点在以簇头无人机为中心，半径为 R_1 的圆形区域内移动，簇成员以各自参考点为中心，在半径为 R_2 的圆形区域内移动，各参考点的移动半径与簇内成员的移动半径之和不超过最大通信距离 $D_{\max}$，即满足 $R_1+R_2\le D_{\max}$。

(2)干扰模型：干扰机使用马尔可夫干扰模式，它在信道转换期间遵循马尔可夫性质，从信道 l_i^{t-1} 转移到信道 $[l_1^t,l_2^t,...,l_L^t]$ 的概率记为 $\mathbf{pr}^t\triangleq[pr_1^t,pr_2^t,...,pr_L^t]$，$\mathbf{pr}^t$ 初始化为 $[0,1]$ 的均匀分布，其中 L 为信道总数。记操作 A：以均等的概率在 $[1,2,...,L]$ 中随机抽取一个信道 l_{random}^t，设 $\text{Sum}_{pr}=\sum_i^L pr_i^t$，从信道 l_i^{t-1} 转移到信道 l_{random}^t 的概率加上 $\text{Sum}_{pr}/\text{div}$，其中 div 为一个倍数缩放因子。操作 B：持续生成 $[0,1]$ 之间的随机数，若该随机数大于 pr_{continue}，则执行一次操作 A，直至生成的随机数不大于 pr_{continue} 为止。操作 C：将 $\mathbf{pr}^t$ 归一化。训练任务利用操作 ABC 的组合生成，以获得多样的元强化学习任务。

(3)通信模型：考虑将时间划分为 T 个时隙，每个时隙的持续时间为 dt，功率增益由慢衰落和快衰落组成，设 $f_{i,m}^{t,l}$ 表示在时隙 t 信道 l 上快衰落增益，$\mathcal{N}$ 为标准正态分布，簇头无人机 m 和簇成员 i 之间的 $f_{i,m}^{t,l}$ 可以表示为:

$$f_{i,m}^{t,l}=20\times\log(\frac{1}{\sqrt{2}}\|\mathcal{N}(1,0)+j\times\mathcal{N}(1,0)\|)$$

慢衰落表示为 $PL_{i,m}^{t,l}=A\times\text{dis}(i,m)^{-\zeta}$，其中 A 为参考距离 1m 处的路径损耗常量，ζ 为路径损耗分量的衰减指数，$\text{dis}(i,m)$ 表示无人机 i 和无人机 m 之间的距离。在时隙 t，CH m 与 CM i 在第 l 个信道上通信的信道功率增益可以表示为 $g_{i,m}^{t,l}=f_{i,m}^{t,l}PL_{i,m}^{t,l}$。定义干扰机 i 在时隙 t 的链路增益信息为 $\overline{g}_i^t$，传输功率为 $\overline{p}_i^t$，干扰机信道选择为 $\overline{\mathbf{L}}^t\triangleq[\overline{l}_1^t,...,\overline{l}_I^t]$。因此，在时段 t，CH m 和 CM i 之间的信道 l 的干扰可以表示为:

$$V_{i,m}^{t,l}=\sum_{q=1}^{M}\sum_{i=1}^{I}1_{\{l_{q,i}^t=l,q\neq m\}}p_{q,i}^t g_{q,i}^{t,l}+\sum_{n=1}^{N}1_{\{\overline{l}_n=l\}}\overline{p}_n^t\overline{g}_n^{t,l}+\sum_{k=1}^{I}1_{\{l_{m,k}^t=l,k\neq i\}}p_{m,k}^t g_{m,k}^{t,l}$$

令σ^2为高斯加性白噪声的功率，B为每个信道的带宽。则簇头无人机m与簇成员i在时隙t信道l上的数据传输速率为:

$$R_{i,m}^{t,l}=B\log_2\left(1+\frac{p_{i,m}^{t}g_{i,m}^{t,l}}{\sigma^2+V_{i,m}^{t,l}}\right)$$

进一步地，步骤 1 中所述的基于元强化学习的无人集群抗干扰通信技术具体将优化问题建模为一个部分可观测的马尔可夫决策过程（POMDP），智能体m在时隙t的观测值可以定义为$o_m^t\triangleq[\mathbf{G}_m^t,\overline{\mathbf{L}}^t]$，其中$\mathbf{G}_m^t\triangleq[\mathbf{G}_{m,1}^t,...,\mathbf{G}_{m,I}^t]$表示在时隙$t$中，簇头无人机$m$与其簇内所有成员在所有通信信道上的信道功率增益，智能体的动作表示为$a_m^t\triangleq[\mathbf{L}_m^t,\mathbf{P}_m^t]$。其中，$\mathbf{L}_m^t$表示传输信道的选择，$\mathbf{P}_m^t$表示发射功率，智能体$m$在$o_m^t$的基础上选择动作$a_m^t$的策略，可以表示为$\pi_m(a_m^t\mid o_m^t)\triangleq\Pr(a_m^t\mid o_m^t)$，$\Pr$表示概率。设传输能量损耗$E_m^t=\sum_{i=1}^{I}\sum_{l=1}^{L}1\{l_{m,i}^t=l\}p_{m,i}^tT_{m,i}^{t,l}$和跳频开销$F_m^t=\sum_{i=1}^{I}\sum_{l=1}^{L}1\{l_{m,i}^t\neq l_{m,i}^{t-1}\}W$，$\beta$调节$E_{m,i}^t$和$F_m^t$的比重，$W$为单位跳频成本。考虑到干扰机在数据传输阶段$T_{Rx}$内可能切换干扰信道，将传输时间$T_{Rx}$分为两个时间段$T_1$和$T_2$，满足$T_1+T_2=T_{Rx}$.假设每个时隙传输的数据包大小为$K$，则在时隙$t$中，簇头无人机$m$与其簇内成员$i$在信道$l$上实际传输时间为：

$$T_{m,i}^{t,l}=\begin{cases}\dfrac{K}{R1_{m,i}^{t,l}} & ,若\dfrac{K}{R1_{m,i}^{t,l}}<T_1,\\ T_{Rx} & ,其他,\\ T_1+\dfrac{K-R1_{m,i}^{t,l}T_1}{R2_{m,i}^{t,l}} & ,若\dfrac{K}{R1_{m,i}^{t,l}}>T_1且\dfrac{K-R1_{m,i}^{t,l}T_1}{R2_{m,i}^{t,l}}<T_2,\end{cases}$$

其中，$R1_{m,i}^{t,l}$和$R2_{m,i}^{t,l}$分别表示时隙t中T_1和T_2时间段内在信道l上簇头无人机m与其簇内成员i的数据传输速率，设辅助变量$Z_m^t=\sum_{i=1}^{I}1\{R1_{m,i}^{t,l}T_1+R2_{m,i}^{t,l}T_2>K\}$，则智能体$m$通过执行动作$a_m^t$会获得的本地奖励$r_m^t=Z_m^t\delta-[\beta E_m^t+(1-\beta)F_m^t]$，$\delta$为数据传输成功奖励。

步骤 2 和步骤 4 中经验池和采样器将任务Γ中的一个转移$c_m^{\Gamma,t}=(o_m^t,o_m^{t+1},a_m^t,r_m^t)$存入经验池$\mathcal{D}_m$和上下文采样池$\mathcal{C}_m$，每个簇头$m$的经验池和上下文采样池相互独立。

进一步地，步骤 3 中描述获得相应动作的过程为：基于 DDQN 算法新增元强化学习框架，从上下文采样器中抽样多组数据$\mathbf{c}^{\Gamma}$作为上下文任务推理编码器的输入，将隐藏状态h^{t-1}和环境当前观测值o^t作为长短期记忆网络的输入，将从上下文任务推理编码器得到的向量z、长短期记忆网络(LSTM)的输出h^t、环境当前观测值连接作为 DDQN 的输入。具体的，用推理网络$q_\phi(z\mid\mathbf{c}^{\Gamma})$用于估计后验分布$p(z\mid\mathbf{c}^{\Gamma})$，由马尔可夫决策过程（MDP）的置换不变性将$q_\Phi(z\mid c_{1:N})$建模为：

$$q_\phi(z \mid c_{1:N}) \propto \prod_{n=1}^{N} \Psi_\phi(z \mid c_n)$$

其中 $\Psi_\phi(z \mid c_n) = \mathrm{N}\left(f_\phi^\mu(c_n), f_\phi^\sigma(c_n)\right)$，它是先由 $f_\phi^\mu(c_n)$ 和 $f_\phi^\sigma(c_n)$ 的神经网络估计初均值 μ 和方差 σ，然后再从正态分布中进行随机抽样得来。LSTM 用于对序列信息进行处理，在遗忘门中，$f^t = \sigma(W_f[h^{t-1}, o^t] + b_f)$ ，其中的 h^{t-1} 为 $t-1$ 时刻的隐藏状态，o^t 为 t 时刻的状态。

进一步地，步骤 5 中描述的更新方法为：令 $G_t = r^t + \gamma r^{t+1} + ... + \gamma^k r^{t+k}$，在 DDQN 评估网络中 $Q_\pi(o^t, a^t; \mathbf{w}^t) = E[G_t \mid o^t, a^t; \mathbf{w}^t]$ 用于评估在网络参数为 $\mathbf{w}^t$ 时当前观测 o^t 和动作 a^t 的 Q_{eval} 值，目标网络 $\hat{Q}_\pi$ 用于评估下个时刻的 Q_{target} 值，定义损失函数 $\mathcal{L} = (r^t + \gamma \tilde{Q}_{\text{target}} - \tilde{Q}_{\text{eval}})^2$，其中 $\tilde{Q}_{\text{target}}$ 和 $\tilde{Q}_{\text{eval}}$ 为当前观测下最大的 Q 值，即选择动作的 Q 值，Q_π 的参数 $\mathbf{w}$ 使用 Adam 算法进行优化，$\hat{Q}_\pi$ 的参数每隔一段时间从 Q_π 参数复制。

进一步地，步骤 6 中描述的元测试阶段具体过程为：利用预先训练好的模型参数在元测试集上进行测试，在测试过程对模型参数进行微调，测试其快速适应新的环境的能力和最终收敛效果。

下面结合附图及具体实施例对本发明做进一步详细说明。

实施例

本发明地一个实施例具体描述如下，采用 python 仿真，参数设定不影响一般性，将所述方法进行比较的方法有：（1）基于上下文感知的 DDQN 元学习抗干扰算法；（2）DDQN 抗干扰算法；(3) 执行随机动作的元学习抗干扰算法。

训练和测试任务根据干扰模型所述，按照 abcd 操作共生成 300 个任务，选择每种操作生成的概率相同：（a）执行操作 C。（b）执行操作 A，再执行操作 C，其中 div=1。（c）执行操作 B，再执行操作 C，其中 div=2，$pr_{\text{continue}} = 0.7$。（d）执行操作 A，然后执行操作 B，最后执行操作 C，其中 div_A=2，div_B=3。前 200 个任务用于训练，后 100 个任务用于测试。

在元训练阶段，所有无人机被限制在长为 500m、宽为 250m，高度限制在 60m 至 120m 的范围内移动，簇内成员的移动轨迹遵循参考点组移动模型，训练的最大步长设置为 200，每 10 步更新一次网络参数。在元测试阶段，设置最大步长为 100，使用更小的批量大小 $\text{batch} = 16$，更长的参数更新间隔 40，表 1 中列出了其他的仿真参数。

如图 1 所示，在元训练阶段，采用元强化学习方案能够稳健地增强模型性能，在模型达到收敛时往往能够达到更好的效果，这是因为基于上下文任务推理的编码器和长短期记忆网络有出色的任务推断能力和记忆保持能力，编码器通过对任务上下文的深入理解，能够准确地捕捉到不同任务之间的共性和差异，从而为后续的策略选择提供有力的支持，而 LSTM 则凭

借其强大的记忆功能，有效地记住了在历史交互过程中积累的经验和知识，使得模型在面对类似情况时能够迅速做出正确的反应。

如图 2 所示，在元测试阶段，从测试集随机抽样任务，并在新任务上测试训练好的模型性能，使用基于上下文感知的 DDQN 元学习抗干扰算法收敛速度明显优于其他方法，这是因为元强化学习通过在多个任务上学习，使得模型能够更快速地学习适应新任务的能力，这种泛化能力使得模型在面对新任务时表现更加出色，并且模型学习到了任务之间的通用特征和模式，而非仅仅记忆特定任务的数据，这种学习方式使得模型更具泛化能力，能够更好地适应各种任务。

如图 3～图 5 所示，基于上下文感知的 DDQN 元学习抗干扰算法的模型通信成功奖励更高，在保证通信质量的同时使用了更小的功率进行通信，使得整个网络的整体寿命得以延长，除此之外，使用了元强化学习的算法能够显著的降低跳频开销，在相同的跳频开销下能够达到更高的通信成功率。

表 1 仿真参数

系统参数			
参数名	值	参数名	值
簇头 M 数	3	簇成员 I 数	2
干扰机数 N	2	信道数 L	4
时隙	1.18 s	传输时间	0.98s
无人机传输功率	[36, 33, 30, 27]	干扰时间	2.28s
天线增益	3 dBi	干扰机传输功率	30 dBm
加性白高斯噪声	-144 dBm	相关加权系数 μ	0.8
参考点 RP 运动半径 R1	99 m	最大通信范围	100 m
带宽 B	1.8 MHz	CMs 运动半径 R2	1 m
飞行高度	60m~120m	数据包大小 K	1MB
区域最大长	500	区域最大宽	250m
训练参数			
参数名	值	参数名	值
回合数	2000	最大步数	200
学习率 α	0.002	权重因子 β	0.5
经验池容量	200×100	批量大小	32
折扣因子 γ	0.9	成功奖励 δ	1
激活函数	ReLu	优化器	Adam

以上显示和描述了本发明的基本原理、主要特征及优点。本行业的技术人员应该了解，本发明不受上述实施例的限制，上述实施例和说明书中描述的只是说明本发明的原理，在不

脱离本发明精神和范围的前提下，本发明还会有各种变化和改进，这些变化和改进都落入要求保护的本发明范围内。本发明要求保护范围由所附的权利要求书及其等效物界定。

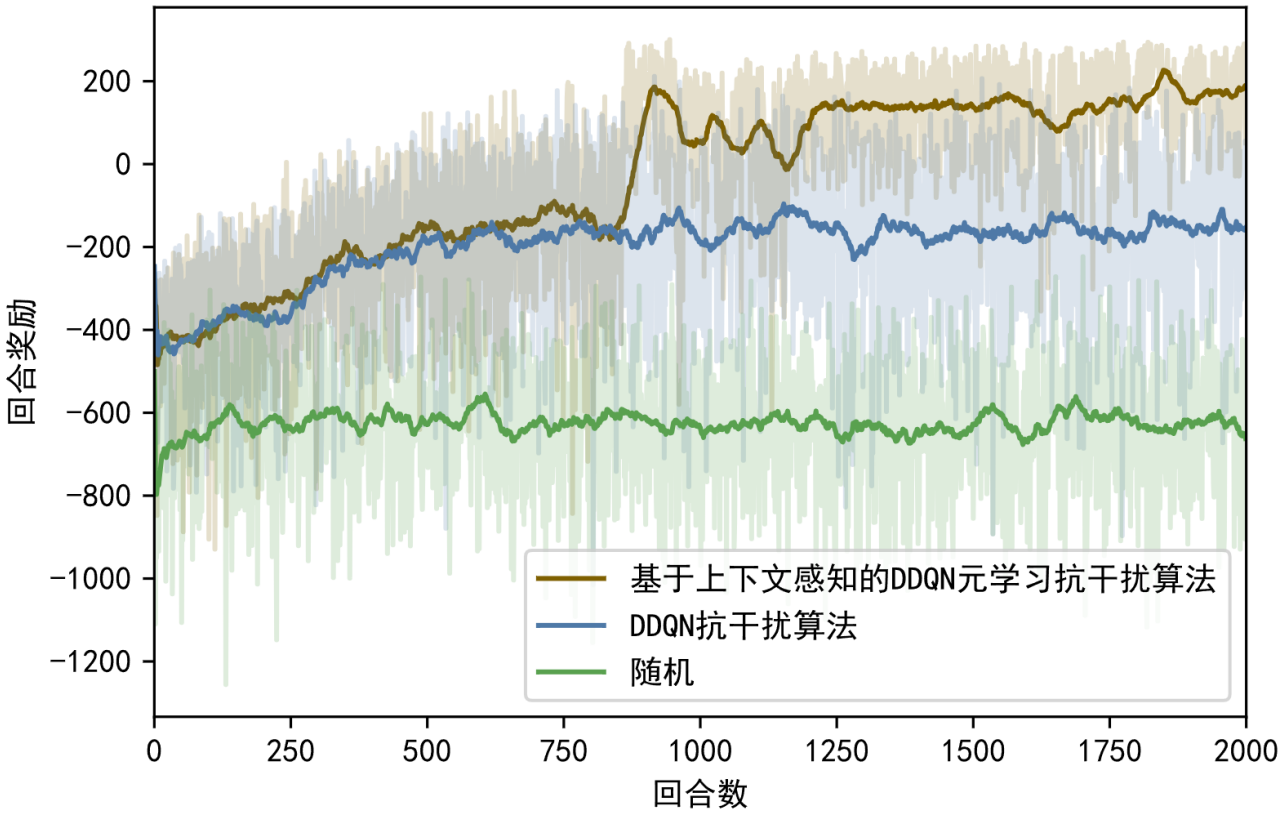

200
0
−200
−400
−600
−800
−1000
−1200
回合奖励
0
250
500
750
1000
1250
1500
1750
2000
回合数
基于上下文感知的DDQN元学习抗干扰算法
DDQN抗干扰算法
随机

图 1

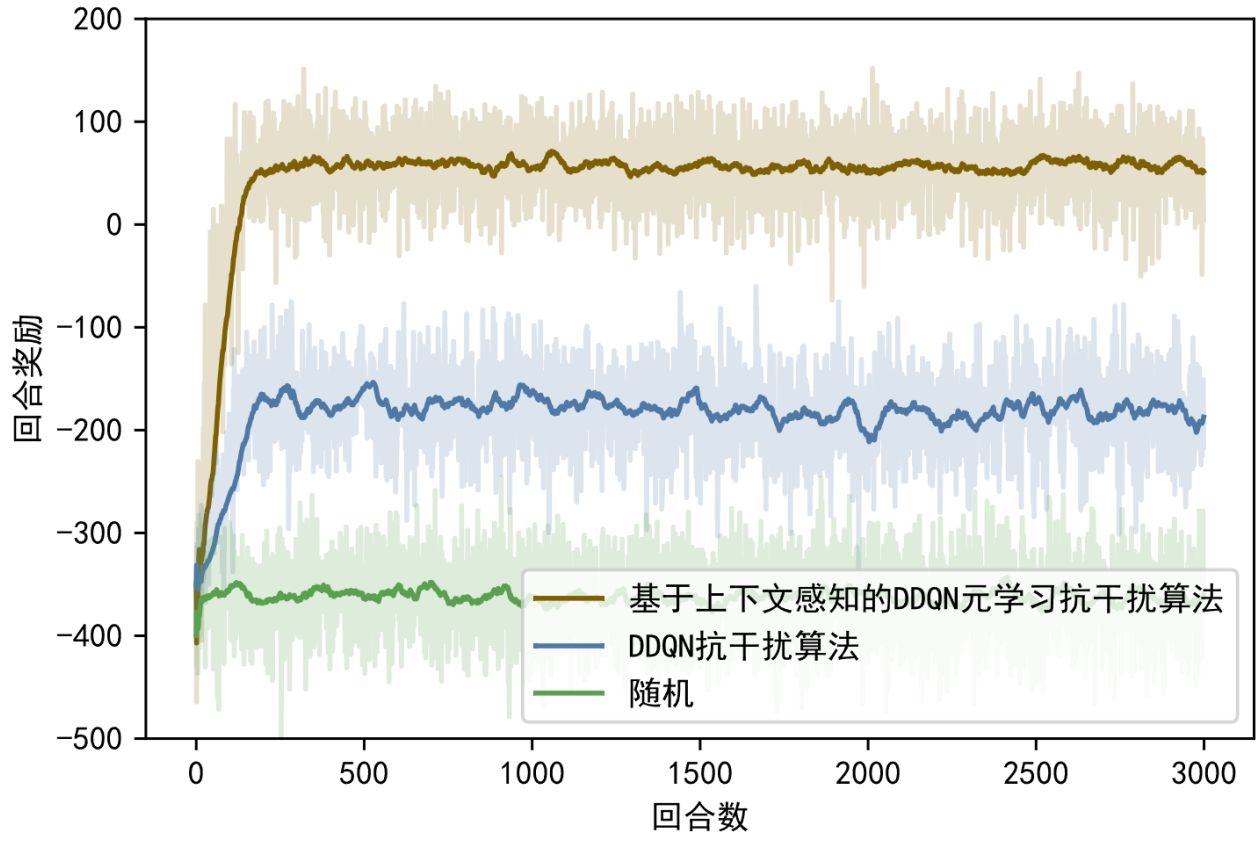

200
100
0
−100
−200
−300
−400
−500
回合奖励
0
500
1000
1500
2000
2500
3000
回合数
基于上下文感知的DDQN元学习抗干扰算法
DDQN抗干扰算法
随机

图 2

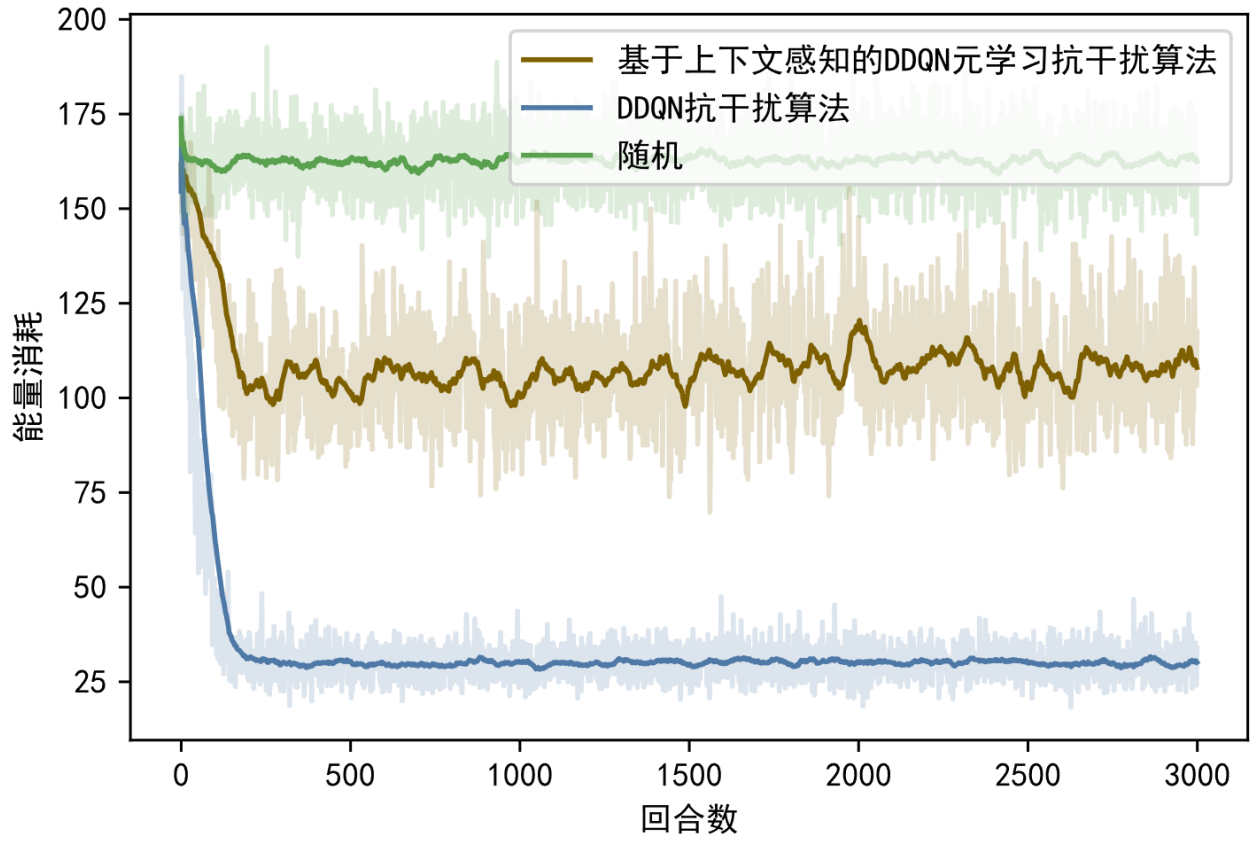

基于上下文感知的DDQN元学习抗干扰算法
DDQN抗干扰算法
随机
能量消耗
回合数
200
175
150
125
100
75
50
25
0
500
1000
1500
2000
2500
3000

图 3

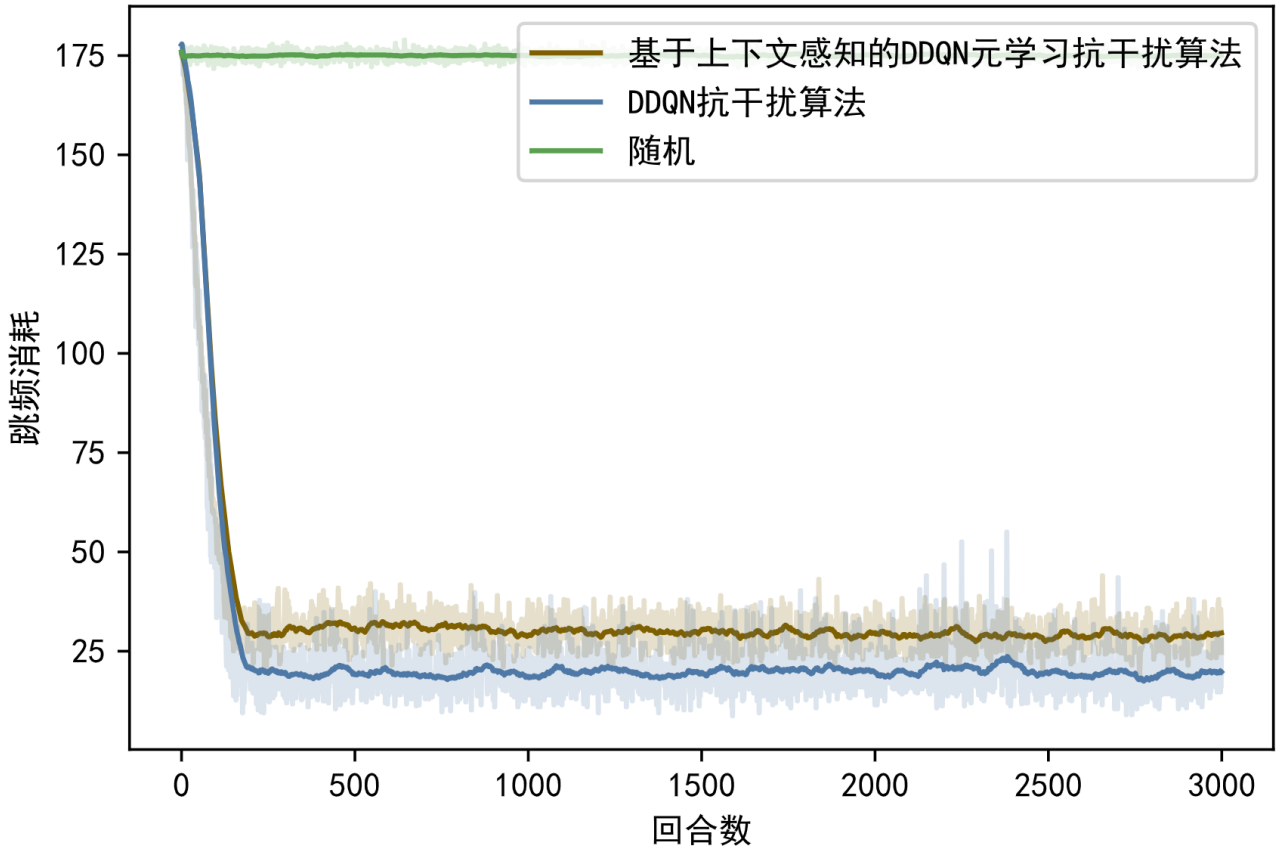

基于上下文感知的DDQN元学习抗干扰算法
DDQN抗干扰算法
随机
跳频消耗
回合数
175
150
125
100
75
50
25
0
500
1000
1500
2000
2500
3000

图 4

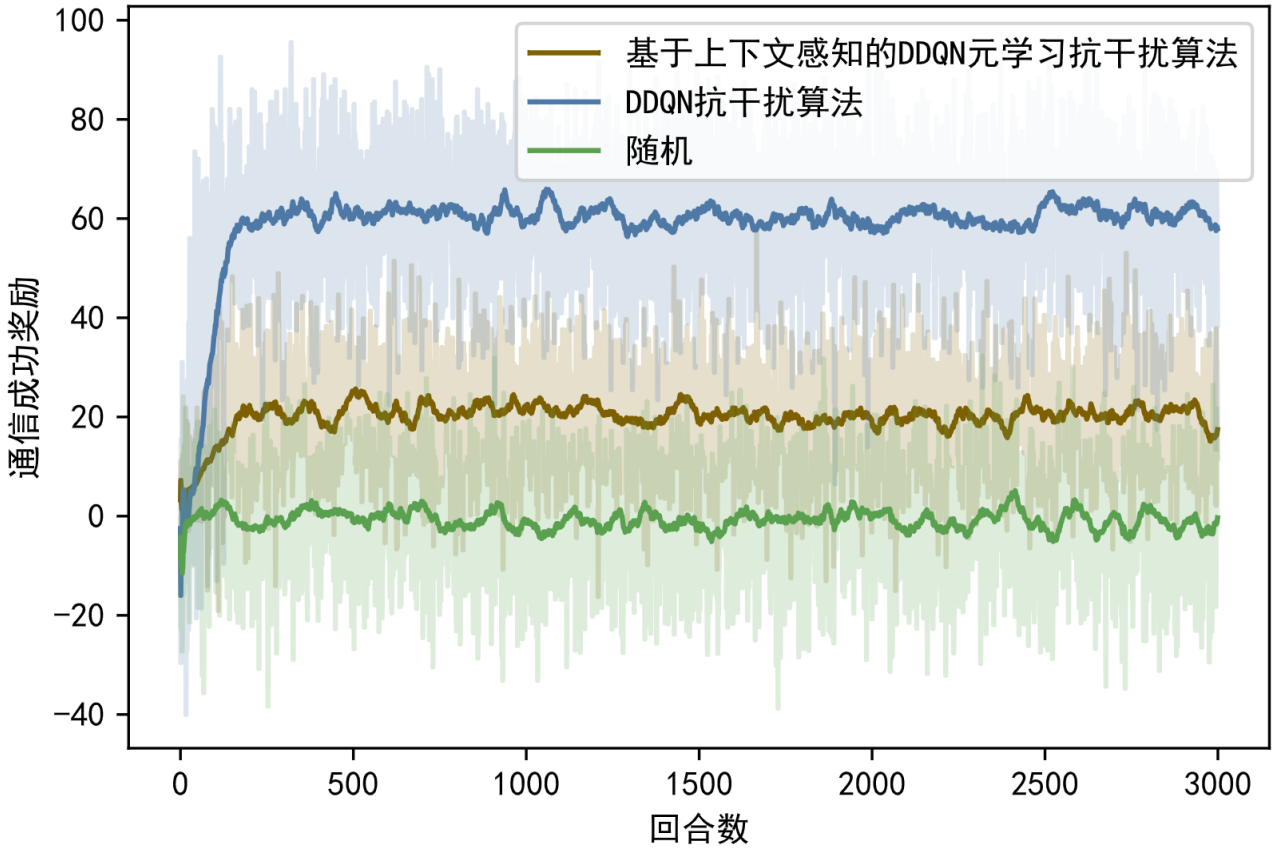

基于上下文感知的DDQN元学习抗干扰算法
DDQN抗干扰算法
随机
通信成功奖励
回合数
100
80
60
40
20
0
−20
−40
0
500
1000
1500
2000
2500
3000

图 5